Kalpa Publications in Computing

Volume 22, 2025, Pages 894–908

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

A Domain Knowledge-Enhanced Large Vision-Language Model for Construction Site Safety Monitoring

Chak-Fu Chan[1], Xiaowen Guo[2], Peter Kok-Yiu Wong[3], Jolly Pui-Ching Chan[4], Jack C.P. Cheng[5], Pak-Him Leung[6] and Xingyu Tao[7]

1) M.Phil. Candidate, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR. Email: cfchanay@connect.ust.hk
2) Ph.D. Candidate, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR. Email: xguoaw@connect.ust.hk
3) Post-doctoral Fellow, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR. Email: cekywong@ust.hk
4) Electrical and Mechanical Engineer, Drainage Services Department, the Government of the Hong Kong Special Administrative Region, Wanchai, Hong Kong Island, Hong Kong, SAR. Email: pcchan02@dsd.gov.hk
5) Professor and Associate Head, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR. Email: cejcheng@ust.hk
6) Co-Founder, AutoSafe Ltd. Email: issacleung@autosafe.ai
7) Research Assistant Professor, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR. Email: xtaoab@connect.ust.hk

Abstract: To address the industry-wide and policy-driven requirements toward construction site safety monitoring, this paper develops a virtual assistant agent based on a large vision-language model (VLM), integrated into on-site surveillance camera system for real-time identification and alerting of unsafe worker behaviors. First, we designed a semi-automatic image-text labeling pipeline, employing in-context learning to enhance data annotation efficiency. Then, we established a two-stage curriculum learning paradigm to deeply embed construction domain knowledge into the VLM, which is eventually embedded into a real-time video analytical engine for safety compliance inspection and interactive visual question answering. The system has been deployed on a real construction site, with around 90% accuracy in identifying violations of work-at-height safety regulations.

Keywords: Construction Site Safety Monitoring, Data-Efficient Fine-tuning Strategy, Domain-Tailored Large Vision-Language Model, Multi-modal Safety Compliance Checking, Virtual Construction Safety Assistant

1. INTRODUCTION

The construction sector is recognized as one of the most dangerous sectors worldwide, with a history of high accident and casualty rates. From 2013 to 2019, the construction sector in Hong Kong has consistently recorded the highest death rates compared to the other 14 primary industry sectors (Labour Department, 2018, 2019). Similarly, in 2019, the United States reported that approximately 20% of work-related deaths took place at construction sites (Occupational Safety and Health Administration, 2019). Many of the casualties are related to non-compliance with construction safety

rules. Construction sites are characterized by safety risks that arise from complex interactions between numerous workers and machinery. To prevent safety risks and serious injuries, it is crucial to detect and analyze any unsafe practices among workers in real-time. This can involve monitoring workers' activity on-site and ensuring they're complying with the safety rules. Traditional site safety monitoring mainly involves regular on-site safety inspection by management personnel and safety inspectors, which is labor-intensive and easily leads to overlooking unsafe behavior. In recent years, analyzing construction images or videos by computer vision (CV)-based deep learning (DL) methods have been widely studied for identifying unsafe construction objects and behavior (Cheng et al., 2022), issuing real-time alerts or more in-depth safety analyses.

However, existing CV-based safety analysis characterized by Convolutional Neural Network (CNN) models has been faced with several problems. First, developed CV models are specialist models only well-trained for a small subset of detection tasks, thus limited by the narrow domain of knowledge. For example, an object detection model trained to perform decently in PPE detection may fall short in fall or injury detection. Second, the architecture of CNN-based models leads to limited embedded semantic information (e.g. object category and localization), while extracting semantic information (e.g., construction activities, interactions between different construction objects) from construction images is an essential step for further CV-based application in construction management (Paneru & Jeelani, 2021; Y. Wang et al., 2022). More high-level tasks like safety compliance checking require a high-level and comprehensive semantic understanding of the on-site scene.

Other researchers have integrated the developed CV models with some other Natural Language Processing (NLP) techniques like pre-defined knowledge graphs for further compliance checking (Fang et al., 2020; Tang et al., 2020). However, processing information separately with vision and language modules may be inefficient and time-consuming. Moreover, during the process of transferring safety information across different modalities, features not extracted in the vision modality will inevitably be lost and not represented in the language modality.

The recent advancement in Large Language Model (LLM), based on Generative Pre-trained Transformer (GPT), has shown tremendous potential for human-like reasoning and conversation. To enable multi-modal tasks like visual question answering (VQA), Large Vision-Language Model (VLM) is further developed with an encoder-decoder architecture, where the encoder processes visual information and the decoder generates textual representation (Chen et al., 2023). An illustrative example is the GPT-4V developed by OpenAI (OpenAI et al., 2024), which has demonstrated human-like reasoning capabilities in combining natural language, texts, and images into VQA processes.

Enabled by the GPT-based pre-trained large vision and language backbones, VLMs can extract both intricate visual and linguistic features, undergo deep fusion between them and generate detailed image descriptions, and even perform multiple types of instruction-following tasks such as multi-round conversion that require sophisticated visual-semantic understanding and reasoning. The strong capability of these pre-trained VLMs made them a strong candidate in many traditional vision-language tasks like image captioning, visual grounding, and VQA in a zero-shot or few-shot setting without any fine-tuning. Compared with CV-NLP-based methods, VLMs enable deep fusion between vision and language features for more comprehensive safety analysis.

While pre-training of VLMs from a large corpus of image-text pairs can align the visual encoder with the language backbone's word embedding space and achieve all-round performance in general tasks, fine-tuning can be performed on a much smaller but more supervised instruction-following dataset to adapt the model to domain-specific context. However, there remain challenges when adapting a pre-trained VLM to tasks related to construction safety analysis and management via fine-tuning.

1) Fine-tuning of VLMs requires a vast amount of labeled and high-quality data. The construction industry is known for limited open access to structured data for DL training, and currently, there is a lack of public datasets for effectively fine-tuning the VLMs toward diverse construction safety tasks. However, the fine-tuning of LLM may require thousands of images, together with high-quality

instruction-response annotations, which is a resource and labor-consuming task, compared to those of traditional VQA and image-captioning datasets which may just consist of a few words as the response for each image.

2) Similar to CNN-based models, VLMs are mainly pre-trained on internet-crowdsourced images, which have notably different image quality from those captured on real construction sites by surveillance cameras. There is a significant data distribution shift when trying to adapt a VLM for real-site downstream tasks. Moreover, the fine-tuning of VLMs in the construction safety domain depends on both images and high-quality instruction-response sets for learning domain-specific knowledge. Efficient learning of domain knowledge with limited vision and text data remains a critical research gap that needs to be addressed by special fine-tuning strategies.

2. METHOD

To address these fundamental challenges in adapting VLMs for construction safety applications, we propose a comprehensive framework that systematically tackles both the data scarcity issue and the domain adaptation problem. Our methodology, illustrated in Figure 1, encompasses three innovative modules specifically designed to overcome the identified limitations:

(1) A semi-automatic image-text data labeling pipeline to enhance the data preparation efficiency;

(2) A two-stage curriculum learning framework to integrate domain knowledge into the VLM;

(3) A real-time analytical system for automatic incident reporting and interactive VQA.

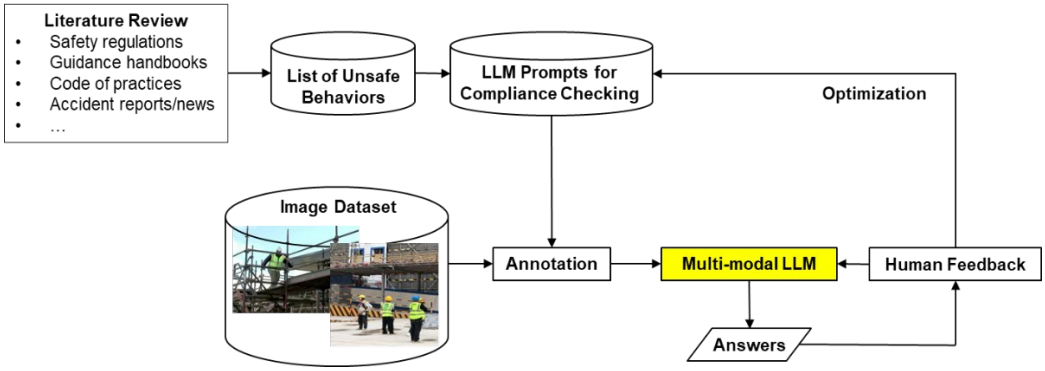

Figure 1: Research Methodology for Developing a VLM-based Safety Monitoring System

2.1 Semi-automatic Data Labeling Pipeline

This section describes the multi-step process to create a high-quality, multi-modal dataset for fine-tuning the VLM for safety compliance and monitoring on construction sites. A semi-automatic pipeline is developed to facilitate image-text data labeling, driven by an in-context learning framework, as illustrated in Figure 2.

.

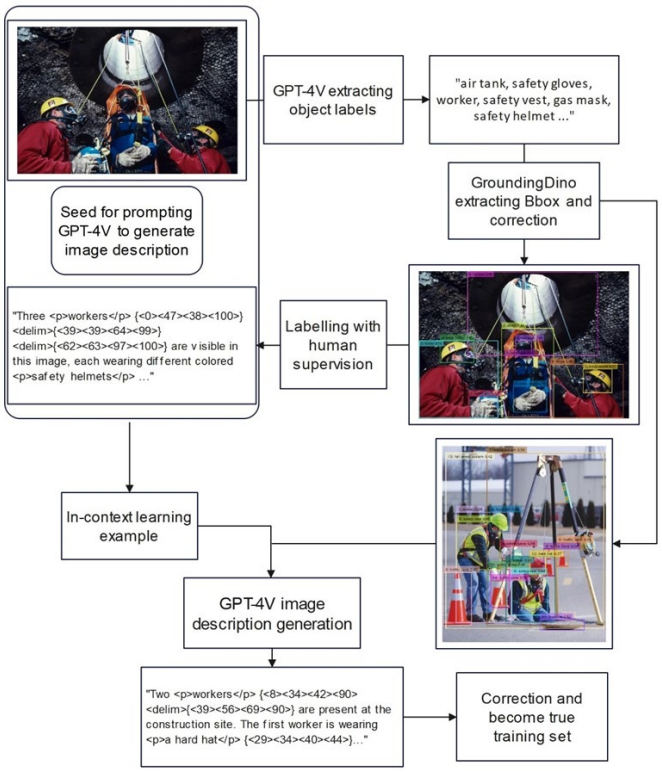

Figure 2: A Self-instructed Pipeline for Semi-automatic Image-text Data Labeling

Specifically, the semi-automatic data labeling pipeline will perform several tasks:

1) Automatic Label and Bounding Box Generation: GPT-4V published by OpenAI is utilized to generate preliminary descriptions (pseudo labels) of each image in batches. Figure 2 shows the assisting pipeline utilizing GPT-4V for effectively extracting candidate object labels or phrases and also a strong open-set object detector named GroundingDINO (S. Liu et al., 2024) for extracting candidate bounding boxes for identified objects.

2) Manual Correction of Pseudo-Labels: The automatically generated pseudo-labels and bounding boxes will be manually corrected to ensure accuracy and reliability.

3) Seed Image Selection: A seed image will be chosen from the images of each scenario (e.g., confined space as shown in Figure 3), and the description of the seed image will be hand-crafted based on the corrected bounding box labels either as grounded captioning (detailed descriptions) or safety compliance analysis examples.

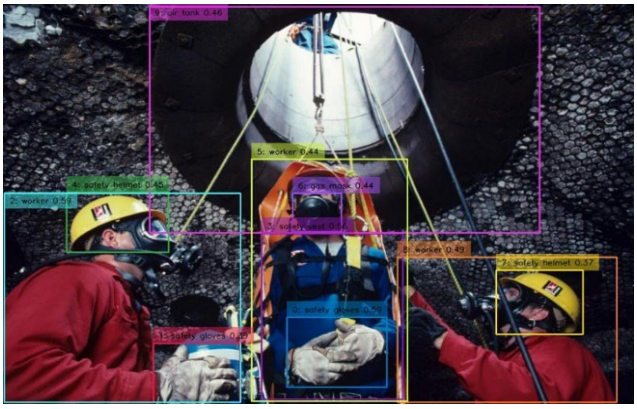

Figure 3: A Seed Image Generated by GPT-4V+GroundingDINO

4) In-context Learning: The annotated seed images will be included in the in-context learning prompt of GPT-4V to generate candidate image descriptions for remaining images. The in-context learning prompt for automatic dataset generation is shown below:

5) Polishing Candidate Descriptions: The candidate image descriptions generated by GPT-4V will be further polished to create more desired responses in the training dataset triplets. These triplets will contain the images, instructions, and the final, polished responses. Lastly, the candidate image description will be further polished to the true desired responses in the training dataset triplets containing images, instructions, and responses. The styles of the instructions across different images will be changed to introduce higher diversity. This diversification of the instructions can help the VLM models learn more robust and generalizable capabilities.

2.2 Two-stage Curriculum Learning Framework for Effective VLM Training

The architecture of CogAgent (Hong et al., 2023) is chosen as our baseline model, as illustrated in Figure 4, CogAgent is built on a pre-trained Vision Language Model (VLM), specifically the CogVLM 17B, which is an open-source and state-of-the-art large vision language model.

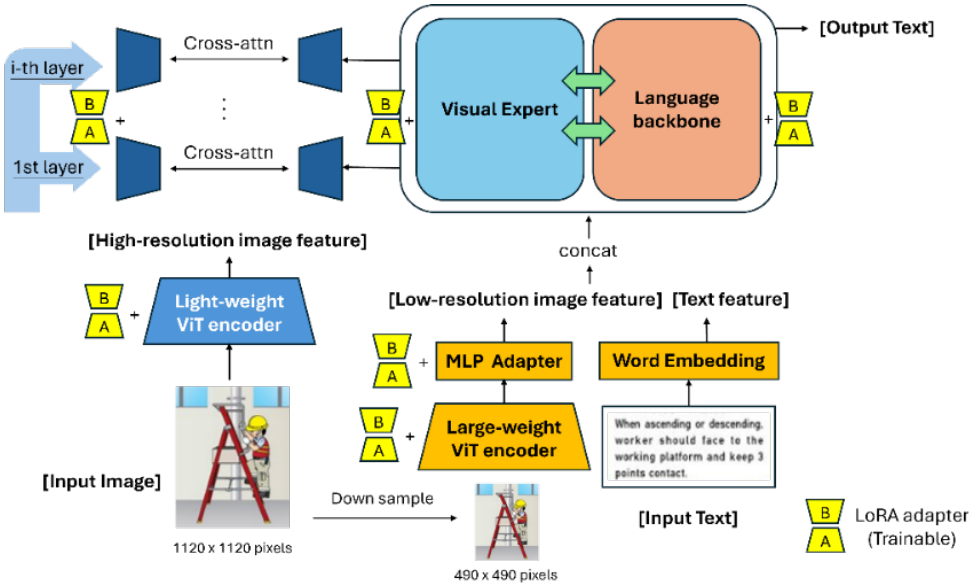

Figure 4: Architecture of the Proposed VLM Architecture (Revised from CogAgent (Hong et al., 2023))

It uses EVA2-CLIP-E as the encoder for low-resolution images (490×490 pixels), and an MLP adapter to map its output into the feature space of the visual-language decoder. The visual-language decoder consists of both vision experts and language backbone for effective deep fusion between image feature and language feature. The decoder processes a combined input of the low-resolution image feature sequence and text feature sequence, and autoregressively outputs the target text. However, the original CogVLM can only accommodate images of relatively low resolution (224 or 490), which is insufficient for screen resolution of computers or smart devices is typically 720p (1280 × 720 pixels) or higher. To address this, a high-resolution cross-module is introduced in their CogAgent model architecture, which not only maintains efficiency with high-resolution images but also offers flexible adaptability to a variety of visual-language model architectures. The high-resolution cross-module acts as a new branch for higher-resolution input, accepting images of size 1120 × 1120 pixels. Unlike the original low-resolution input branch, the high-resolution cross-module adopts a much smaller pre-trained vision encoder (visual encoder of EVA2-CLIP-L in our implementation, 0.30B parameters), and uses cross-attention of a small hidden size to fuse high-resolution image features with every layer of VLM decoder, thus reducing the computational cost.

In terms of the attention procedure, each layer's attention module is formulated in Equations (1) and (2):

$$X_i' = MSA(layernorm(X_{in}^i)) + X_{in}^i \quad (1)$$

$$X_{out}^i = MCA(layernorm(X_i'), X_{hi}) + X_i' \quad (2)$$

MSA and MCA respectively represent multi-head self-attention with visual expert and multi-head cross-attention. The cross-attention with high-resolution images can be perceived as a complement to the features of low-resolution images, thereby effectively utilizing the previous pre-trained model in low resolution. CogAgent exhibits exceptional performance on benchmarks that assess referring expression comprehension (REC), such as RefCOCO, RefCOCO+, and RefCOCOg, which is

comparable to grounding specialists like GroundingDINO.

Current studies adopted a staged learning scheme to train VLMs more effectively. For instance, MiniGPT-v2 was trained with three stages (pretraining, multi-task training and multi-modal instruction tuning) for better visual-textual feature alignment (Chen et al., 2023). Inspired by such principle, a two-stage curriculum learning strategy is proposed to equip our VLM with domain knowledge and capabilities in multiple safety-related tasks:

1) The first stage involves pre-training of our model on construction site images with grounded captioning labels, for learning construction-specific domain knowledge given the labeled construction site objects.

2) The second stage focuses on fine-tuning the VLM toward site-specific safety rules to gain task-specialized performance in safety compliance inspection from images.

This two-stage paradigm ensures that our VLM not only understands the nuances of the language but also adapts to the specific requirements of the task. The dataset decreases in volume but increases in supervision as the training stage progresses. The final stage of instruction tuning aimed to enhance the instruction following and conversation ability of CogVLM.

We also designed a curriculum learning scheme to align the base CogAgent model with advanced safety inspection ability. The first stage of the training mainly involved training the model on grounded captioning datasets of construction images to equip it with construction domain knowledge. Afterwards, the second stage of the training aimed to enhance the inspection ability of the model by learning from specific safety rule compliance/violation cases.

A list of essential work-at-height behaviors from construction documents, including the Safety Manual established by Drainage Services Department (2018), the Work-at-Height Safety Handbook published by Development Bureau and Construction Industry Council (2019) of Hong Kong are reviewed to extract some key safety rules for working at height. The list of work-at-height behaviors are:

a. Powered-Operated Elevating Working Platforms (PEWP)
   i. Wear full body safety harness with its lanyard anchored to a specified anchorage point

b. Metal Scaffolds
   i. When erecting, altering dismantling of scaffolds or it is impracticable to erect a safe working platform or provide safe access and egress, the use of full body safety harness attached to secure anchorage point or an independent lifeline is required

c. Light-Duty Working Platform/ Mobile Working Platform
   i. When ascending or descending, worker should face to the working platform and keep 3 points contact
   ii. Only three types of platforms are allowed to carry out work-at-height tasks, including hop-up platform, step platform and mobile platform
   iii. The surrounding of working platforms should be kept free from waste and miscellaneous materials

d. Floor Opening/Edge Protection
   i. Provide guard-rails and toe-boards at the floor edge
   ii. Provide secure coverings with warning signs at the floor opening.
   iii. Provide guard-rails, toe boards and warning signs at the floor opening
   iv. Whist installing, alternating or dismantling fall protection facilities at the floor edge, opening and windows, suitable fall arresting system should be provided to workers

2.3 Real-time on-site analytical system

The VLM-based safety compliance monitoring system operates in two different modes: (1)

automatic reporting and (2) interactive chatbot. In Mode 1, the real-time video from each camera is streamed continuously to the computer, where video frames are sampled at regular intervals to be fed into the VLM for inference (10 seconds in our field trials). Then, the VLM automatically generates appropriate answers to the image. On the other hand, Mode 2 can be user-triggered via a simple button to extract the current frame at any time. Based on that frame, safety officers can input any question to prompt the VLM to interactively generate personalized answers.

An alarm mechanism is developed to generate descriptive alert messages to safety officers. As illustrated in Figure 5 (left), in case any safety rule violation is identified, such "Failed" safety compliance is reported via an alarm from the onsite computer, where the safety officers are notified in real-time. A descriptive alert message is then prompted out, concisely summarizing which specific safety rules are violated. In case no violation is observed from an image, as illustrated in Figure 5 (right), it passes the safety compliance checking by our VLM which results in a simple message "No Safety Risk" output without triggering the alarm system.

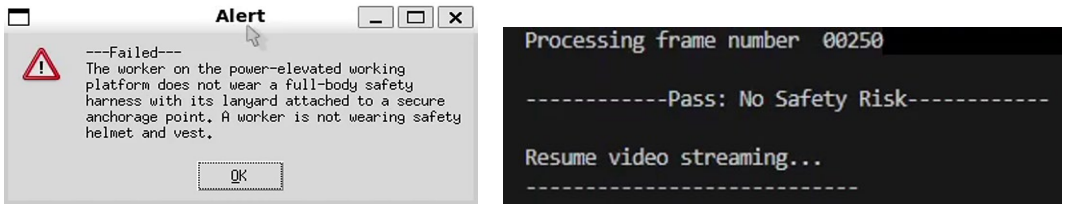

Figure 5: Illustrations of With-Alert and No-Alert Messages Generated by our Alarm System

In addition, an incident logging system is developed to store all the processed images and generated output. These data are properly indexed/labeled, which constitute a well-managed database for record-keeping, which also allows safety officers to query historical incidents and search for associated images for verification. As illustrated in Figure 5, each processed frame is indexed with a number (e.g. "00250" denotes the 250th frame of a video stream), which serves as the key for future queries. All these indexed frames are stored, each of which is then labeled with the final output generated by the VLM, which serve as the value for future queries.

3. RESULTS AND DISCUSSION

3.1 Dataset and Experimental Setup

The developed VLM system is deployed at a real construction site in the Shek Wu Hui Sewage Treatment Works. This site is a secondary sewage treatment plant occupying 9.4 hectares of land and handling 81,000 m^3 of sewage per day produced by a population of 300,000 in Sheung Shui and Fanling Districts.

The dataset for fine-tuning the CogAgent model consists of 1,500 construction site images, after applying augmentation techniques including random cropping, horizontal flipping and color space alternation. The collected images are then annotated with the assistance of the pipeline mentioned in Section 2.1. As shown in Figure 6, the raw images are first processed by the proposed methodology to obtain accurate bounding boxes for objects with precise and diverse textual descriptions. The textual descriptions are based on the list of safety questions summarized. Some features/objects in the images are highly relevant to certain safety rules (e.g. scaffolds, working platforms), thus are specifically highlighted in the annotations to incorporate such knowledge into our VLM.

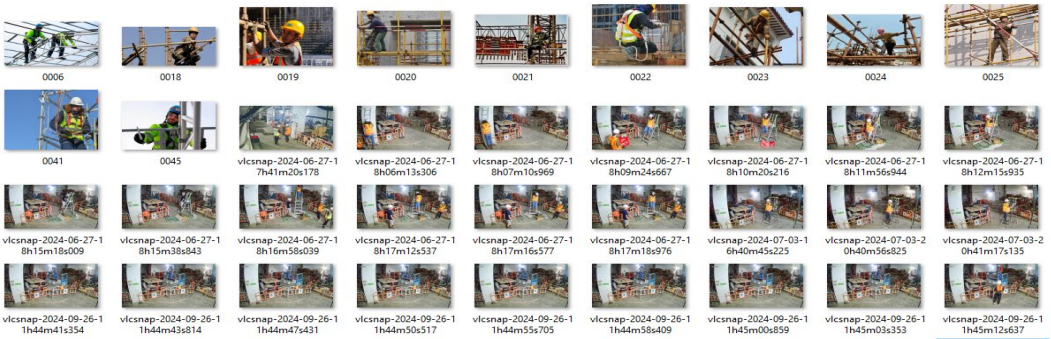

Figure 6: Snapshot of the Collected Image Dataset for VLM Fine-tuning and Testing

The data split is 6:4, i.e. 60% (900 images) for fine-tuning and validation, the remaining 40% (600 images) for testing. The fine-tuning was done with a single Nvidia A40 GPU with 48 GB of available VRAM. For the training configuration, Low-Rank Adaptation (LoRA) is used for Parameter-Efficient Fine-Tuning (PEFT) while all the original weights are frozen. 5 types of LoRA modules, each with a rank of 64, and a LoRA alpha value of 128, are injected into the visual language backbone, the high-resolution cross-attention layers, the large ViT encoder, the MLP adapter and the lightweight ViT encoder respectively. The VLM was trained with a batch size of 2 for 2 epochs for each of the two stages, with a warmup ratio of 0.1, an initial learning rate of 3e-5, and a cosine learning rate scheduler. The lora dropout is set to 0.1 and the weight decay is set to 0.05. After finetuning, the VLM was quantized to 8 bit and deployed in one RTX-4090 GPU with 24 GB of VRAM.

3.2 Quantitative Evaluation

The collected images are fed into our VLM to carry out VQA. The performance of safety compliance checking is evaluated with the protocol summarized in Table 1. In the context of safety compliance checking in this project, two possible outcomes are defined: (1) *Compliance – Pass*, where no safety rule is violated, and (2) *Compliance – Failed*, where a particular safety rule is violated. Therefore:

- True Positive (TP) means that an image does contain violation scenario(s), and the VLM correctly identifies them. Note that the evaluation is counted on per-rule basis rather than per-image basis.
- False Negative (FN) means that an image does contain violation scenario(s), but the VLM cannot identify them and wrongly output "Pass – No Safety Risk".
- False Positive (FP) means that an image has no safety violation, but the VLM wrongly outputs "Failed" for a particular safety rule.
- True Negative (TN) means that an image has no safety violation, and the VLM correctly outputs "Pass – No Safety Risk".

These metrics form the foundation for our two primary evaluation measures, Sensitivity and Specificity, defined in Equations (3) and (4):

- *Sensitivity* denotes the percentage of violation scenarios correctly identified by the VLM.
- *Specificity* denotes the percentage of compliant scenarios correctly passed without any false alert.

Table 1: Proposed Evaluation Metrics for Safety Compliance Checking

Ground-truth \ Output	*Compliance – Failed*	*Compliance – Pass*
Compliance – Failed	*TP*	*FN*

Compliance – Pass	*FP*	*TN*

*The values denote the number of safety rules being checked one-by-one

$$Sensitivity\ (\%) = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity\ (\%) = \frac{TN}{TN + FP} \tag{4}$$

In the context of construction safety monitoring, these metrics carry significant practical implications. Sensitivity measures the system's ability to correctly identify actual safety violations, such as missing PPE or unsafe behaviors at heights. This metric is particularly crucial as missing violations (false negatives) could lead to serious accidents or fatalities. Complementarily, specificity measures the system's accuracy in correctly identifying safe conditions while avoiding false alarms. High specificity is essential for maintaining operational efficiency, as false positives can cause unnecessary work interruptions and diminish trust in the monitoring system.

Our analysis reveals significant improvements through the two-stage curriculum learning framework. The baseline model initially showed sensitivity and specificity rates of 76.3% and 74.3% respectively. Our enhanced framework substantially improved these metrics to 84.7% and 91.7%. Statistical validation through Fisher's Exact Test yielded p-values of 0.0132 for sensitivity and <0.0001 for specificity, confirming the statistical significance of these improvements (both <0.05).

The detailed results in Tables 3 and 4 demonstrate the tangible impact of our approach, showing an increase of 77 correct samples (25 TP + 52 TN), representing approximately 13% of the testing set. This improvement reflects enhanced accuracy in safety compliance identification and validates the effectiveness of our integrated approach for construction safety monitoring.

Table 2: Quantitative Results among the Baseline and Proposed Methods

Method	Sensitivity	Specificity
Baseline	76.3% (71.1% ~ 81.0%)	74.3% (69.0% ~ 79.2%)
Two-stage Learning	**84.7%** (80.1% ~ 88.6%)	**91.7%** (87.9% ~ 94.5%)

*The ranges inside brackets denote the results at 95% Clopper-Pearson confidence intervals

Table 3: Contingency Table for Sensitivity of Safety Compliance Checking

Method	*TP*	*FN*	Total
Baseline	229	71	300
Two-stage Learning	**254**	**46**	300

Table 4: Contingency Table for Specificity of Safety Compliance Checking

Method	*TN*	*FP*	Total
Baseline	223	77	300
Two-stage Learning	**275**	**25**	300

3.3 Qualitative Evaluation

Based on the safety rules listed in Section 2.2, the output generated by our VLM (for compliance "Pass" and "Failed" respectively) are illustrated below.

Case 1: Working on Light-duty Mobile Platform (Safe Ascending/Descending Required)

- "Pass": In Figure 7, the worker ascending/descending the step platform is facing directly to the platform and maintaining three-point contact with it. Hence, the safety compliance checking correctly results in a "Pass".

- "Failed": As shown in Figure 8, the worker is facing away from the step platform, and his hands are not holding the step platform safely. In that cases, the alarm system is triggered with an alert message prompted out, stating the safety rules being violated.

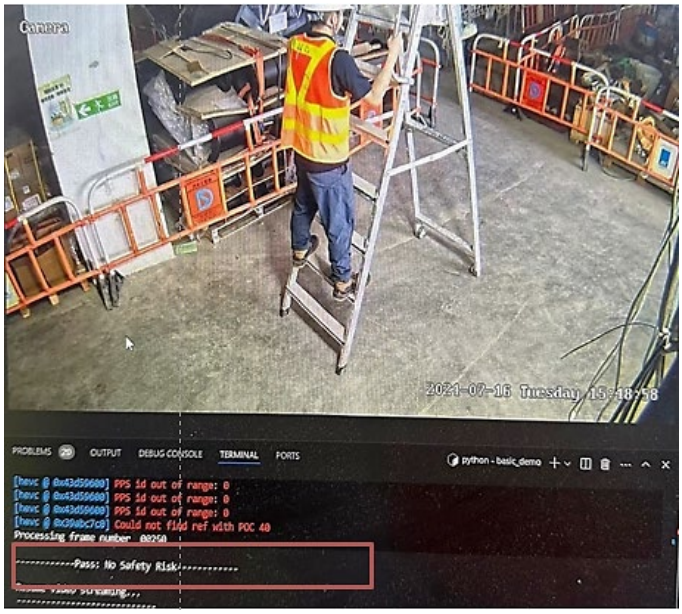

Figure 7: Result of Facing to the Platform and Three-point Contact Compliance ("Pass")

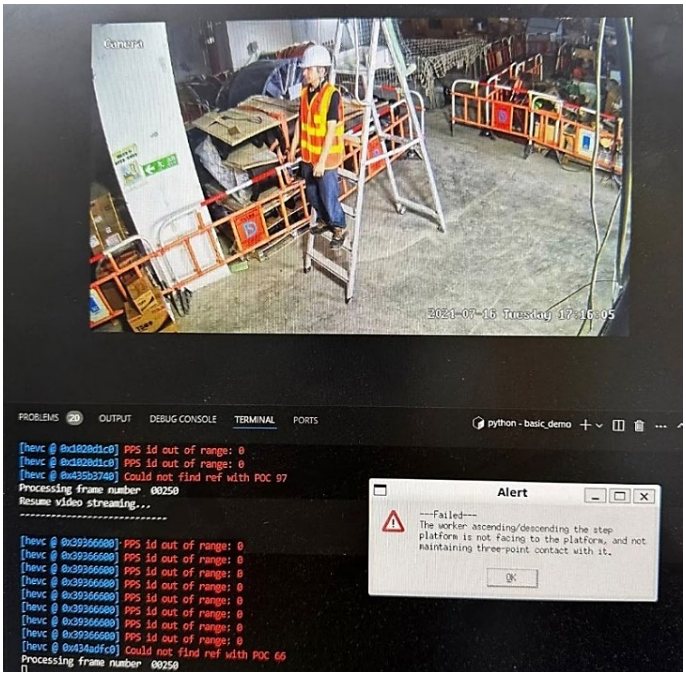

Figure 8: Result of Facing to the Platform and Three-point Contact Compliance ("Failed")

For monitoring worker safety when climbing the mobile platform, our VLM can analyze the orientation of the worker's face and body relative to the mobile platform. However, the identification of three-point contact with the mobile platform is slightly more challenging. The worker's limbs may be partially occluded when climbing the mobile platform (e.g. when facing to the right, his left hand and left leg are occluded respectively by his body and right leg). More systematic prompt engineering may be explored in the future to further enhance the identification accuracy.

Case 2: Working on PEWP (Safety Harness Required)

- "Pass": As shown in Figure 9, the worker standing on top of the PEWP is wearing a safety harness, with its lanyard attached to a secure anchorage point. The VLM simply prints "Pass: No Safety Risk" on the screen to verify his safety compliance, without triggering any alarm.
- "Failed": As shown in Figure 10, the worker on the PEWP is not wearing a safety harness. Our VLM correctly identified such safety violation and generated the corresponding alert message.

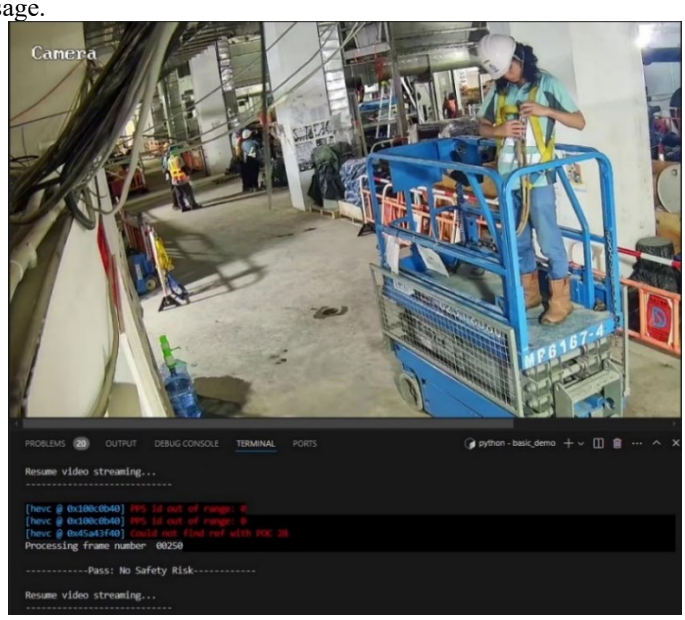

Figure 9: Result of Safety Harness and Anchorage Compliance ("Pass")

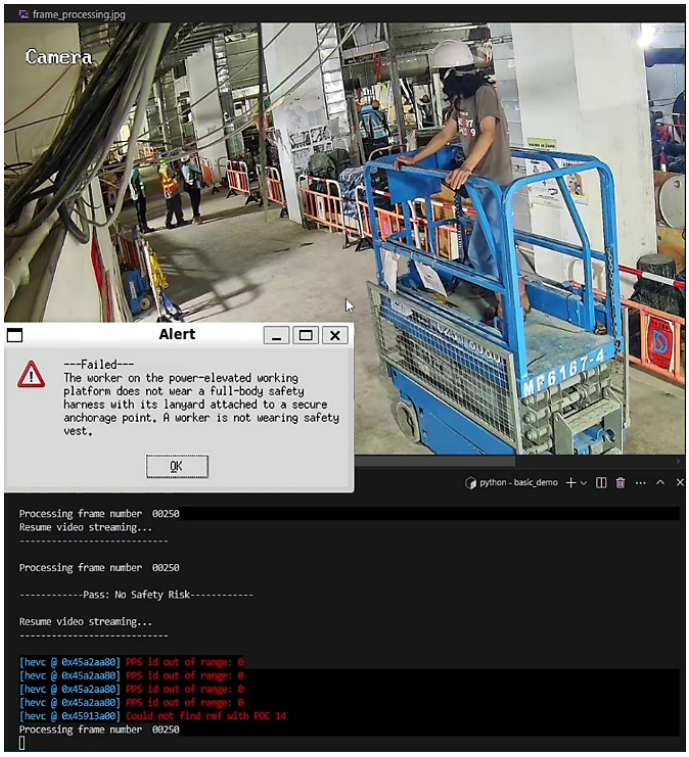

Figure 10: Result of Safety Harness and Anchorage Compliance ("Failed")

Overall, our VLM can accurately identify the presence/absence of safety harness when being worn by a worker. However, it remains challenging to determine whether the harness is attached to a secure anchorage point (in the site trial, attaching to the PEWP is considered safe), due to the very thin rope and small size of the anchor. More crucially, the anchor is usually occluded by the worker or the PEWP itself, making the anchorage detection inaccurate. In the future, more sophisticated strategies of image/video analytics, such as small-object attention mechanism, can be further incorporated into our framework, to enhance the harness detection robustness.

Nevertheless, the preliminary results show that the CogAgent model show some hallucination when transferring the visual information into textual information. This may be due to the frozen visual encoder and shallow alignment method in the model architecture during the model training. Further investigation may be needed on the multi-modal feature alignment capability and contextual awareness toward safety monitoring.

4. CONCLUSION AND FUTURE WORK

This paper addresses critical challenges in adapting VLMs for construction safety monitoring through a comprehensive framework. Our semi-automatic pipeline, combining GPT-4V and GroundingDINO for image captioning, successfully overcame data scarcity. Our two-stage curriculum learning framework demonstrated remarkable effectiveness in domain knowledge integration, achieving 84.7% sensitivity and 91.7% specificity in work-at-height safety compliance checking, substantially improved over the baseline method.

These findings have significant implications for both research and industry. Our framework provides a scalable solution for adapting AI systems to specialized domains with limited data

availability, while the demonstrated success in real-world deployment establishes a practical pathway for automating construction safety monitoring. The high accuracy achieved in safety compliance inspection suggests potential for widespread adoption in construction site management, enabling more proactive and efficient safety protocols. Future research will focus on enhancing the system through architectural modifications for improved multi-modal feature alignment and contextual awareness. The framework's generalizability will be validated across diverse safety rules, and be extended to real-time video analytics for automated safety monitoring practices.

ACKNOWLEDGMENTS

The authors would like to acknowledge the project (No. *DEMP/2023/25*) funded by *Drainage Services Department, the Government of the Hong Kong Special Administrative Region, Wanchai, Hong Kong Island, Hong Kong, SAR*, for providing support to this research. The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., & Elhoseiny, M. (2023). *MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning* (arXiv:2310.09478). arXiv. https://doi.org/10.48550/arXiv.2310.09478

Cheng, JackC. P., Wong, P. K.-Y., Luo, H., Wang, M., & Leung, P. H. (2022). Vision-based monitoring of site safety compliance based on worker *re*-identification and personal protective equipment classification. *Automation in Construction*, *139*, 104312. https://doi.org/10.1016/j.autcon.2022.104312

Fang, W., Ma, L., Love, P. E. D., Luo, H., Ding, L., & Zhou, A. (2020). Knowledge graph for identifying hazards on construction sites: Integrating computer vision with ontology. *Automation in Construction*, *119*, 103310. https://doi.org/10.1016/j.autcon.2020.103310

Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Zhang, Y., Li, J., Xu, B., Dong, Y., Ding, M., & Tang, J. (2023). *CogAgent: A Visual Language Model for GUI Agents* (arXiv:2312.08914). arXiv. https://doi.org/10.48550/arXiv.2312.08914

Labour Department. (2018). *Occupational Safety and Health Statistics 2018*. Occupational Safety and Health. https://www.labour.gov.hk/eng/osh/pdf/Bulletin2017.pdf

Labour Department. (2019). *Occupational Safety and Health Statistics 2019*. Occupational Safety and Health. https://www.labour.gov.hk/eng/osh/pdf/archive/statistics/OSH_Statistics_2019_eng.pdf

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). *Improved Baselines with Visual Instruction Tuning* (arXiv:2310.03744). arXiv. https://doi.org/10.48550/arXiv.2310.03744

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024). *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection* (arXiv:2303.05499). arXiv. https://doi.org/10.48550/arXiv.2303.05499

Occupational Safety and Health Administration. (2019). *Commonly Used Statistics | Occupational Safety and Health Administration*. OSHA Data & Statistics. https://www.osha.gov/data/commonstats

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Paneru, S., & Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities & challenges. *Automation in Construction*, *132*, 103940. https://doi.org/10.1016/j.autcon.2021.103940

Tang, S., Roberts, D., & Golparvar-Fard, M. (2020). Human-object interaction recognition for automatic construction site safety inspection. *Automation in Construction*, *120*, 103356. https://doi.org/10.1016/j.autcon.2020.103356

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., & Tang, J. (2024). *CogVLM: Visual Expert for Pretrained Language Models* (arXiv:2311.03079). arXiv. https://doi.org/10.48550/arXiv.2311.03079

Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M., & Li, H. (2022). Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Advanced Engineering Informatics*, *53*, 101699. https://doi.org/10.1016/j.aei.2022.101699